XSEDE Data Analytics Use Cases

14th Jun 2013

Version 0.3

Table of Contents

A. Document History

Overall Document Authors:

Shawn Strande (SDSC, Point of Contact)

Natasha Balac (SDSC)

Pietro Cicotti (SDSC)

Amit Majumdar (SDSC)

Nick Nystrom (PSC)

Robert Sinkovits (SDSC)

Mahidhar Tatineni (SDSC)

Nicole Wolter (SDSC)

	Version	Date	Changes	Author
XSEDE Analytics Use Cases	0.1	9/18/12	First Version submitted to A&D	Data Analytics Team
	0.2	4/26/2013	Iterating Architects feedback	Data Analytics Team
	0.3	3/16/2013	Significant revisions to first two use cases based on	Data Analytics Team

			feedback	

B. Document Scope

This document is both a user-facing document (publicly accessible) and an internal working document intended to define user needs and use cases that fall under the general umbrella of Data Analytics within the overall activities of XSEDE. The definition of use cases is based on a template from Malan and Bredemeyer[1]. In general it is in keeping with the approaches and philosophy outlined in "Software architecture in practice."[2]

The use cases are presented here using the following format, derived from the Malan and Bredemeyer white paper[1] as follows:

Use Case	Use case identifier and reference number and modification history
Description	Goal to be achieved by use case and sources for requirement
References	References and citations relevant to use case
Actors	List of actors involved in use case
Prerequisites (Dependencies) & Assumptions	Conditions that must be true for use case to be possible Conditions that must be true for use case to terminate successfully
Steps	Interactions between actors and system that are necessary to achieve goal
Variations (optional)	Any variations in the steps of a use case

[1] Malan, R., and D. Bredemeyer. 2001. Functional requirements and use cases. *www.bredemeyer.com/pdf_files/functreq.pdf*

[2] Bass, L., Paul Clements, and Rick Kazman

Quality Attributes	
Non-functional (optional)	List of non-functional requirements that the use case must meet
Issues	List of issues that remain to be resolved

C. Data Analytics Use Cases

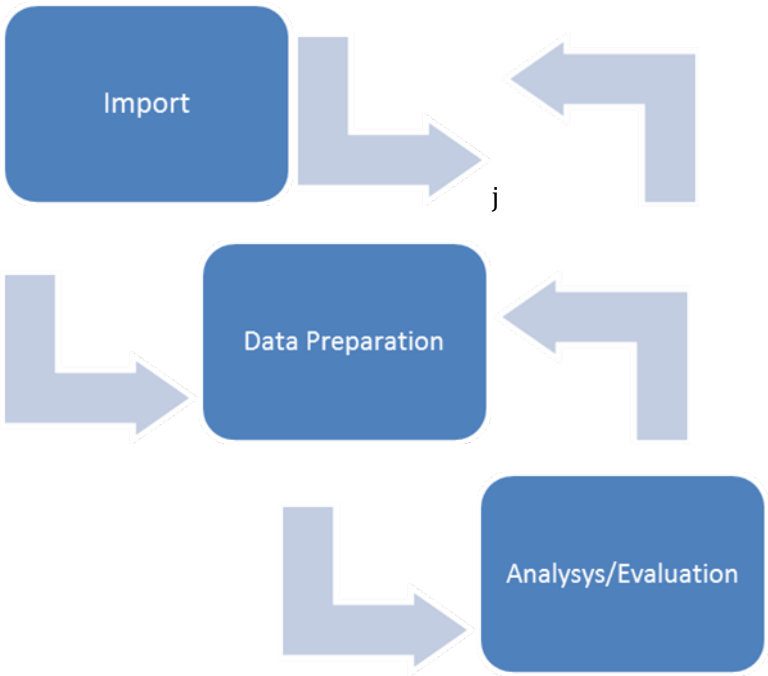
UCDA 1.0	Data Analytics Resources and Information
Description	User obtains information about data analytics in the XSEDE resource portfolio, discovering existing XSEDE analytics resources and accompanying documentation. This is a pure discovery process; no data manipulation or analytics are performed. This information includes: 1) definitions and examples of data analytics in the context of XSEDE; 2) a comprehensive list and user guide for computational resources and software available for data analytics; and 3) support for gaining access to and performing data analytics on XSEDE resources.
References	http://www.xsede.org/resources/overview
Actors	• XSEDE Portal visitor, which in this context is a person interested in using XSEDE for Data Analytics (current or prospective users) • XSEDE Service Providers • XSEDE Portal developers and content managers
Prerequisites (Dependencies) and Assumptions	• XSEDE Portal visitor has Internet access. • XSEDE Portal is available • Information about Data Analytics resource available via XSEDE is accurate • Glossary of terms, or other foundational information • User support in the form of contacts, workshops, training courses
Steps	1. The XSEDE Portal visitor navigates through the Resources area of the XSEDE website to find a “Data Analytics” drop down menu item 2. Upon selection of this item, the Portal returns a categorized list of resource and applications available in XSEDE. 3. From the list of available resources, additional information, such as user guides, allocation instructions, and related detailed information can be accessed. This information may be hosted on

	the XSEDE Portal, Service Provider web resources, or elsewhere. 4. The list of available resources and applications is accompanied by link to a glossary of data analytics terms. 5. The list of available resources includes: references to XSEDE staff or other experts that can assist with data analytics; training workshops; and reference material for data analytics.
Variations (optional)	Variation 1: 1. Portal visitor enters search strings such as "data analytics", or other common terms in the search field of the XSEDE user portal. 2. Portal returns, at a minimum, links to the same information that is available via the drop down in the above case.
Quality Attributes	● **Accuracy**: Information about the application and resources available via XSEDE changes over time and will need to be refreshed as resources come and go ● **Completeness**: All the information necessary to understand the available resources, apply for them, and required for their basic use is available on the XSEDE Portal. ● **Context**: Data analytics resources and information should be defined in terms that are relevant for the XSEDE user community. This may imply that some tools, algorithms, and information, which are part of the larger data analytics space, are not applicable, and therefore not available in XSEDE.
Non-functional (optional)	
Issues	● Data analytics as a field is not well defined. ● The process for creating new items in the XSEDE portal is not defined. ● The process for making updates to existing Portal information is not defined.

UCDA 2.0	Data Preparation
Description	User engages in data preparation, an activity that may include selecting, cleaning, supplementing, integrating, formatting, and modeling data. Users should be able to gather, manipulate, translate and organize their data as needed for data mining, modeling or analysis activities. For example, the computational environment should be configured such that users can: collect data from local and remote resources; store data to archival resources; stage data from archival resources to compute resources; and execute a broad array of data preparation processes, such as cleaning, integration, transforming, reduction summarizing, sampling, etc. XSEDE needs to provide tools for creating metadata, and libraries for reading standard file formats such as NetCDF, HDF5, and others. Data and metadata may be in multiple formats (flat file, relational RDBMS, Hadoop data file system, text, video etc.). Access to data may be local or remote. Some applications may need to access the internet in unconstrained ways. (Note: need to support applications that require Internet scraping, i.e., they make frequent external TCP connections. This may become a property of a particular set of resources enabling these features)
References	• http://www.xsede.org/resources/overview • http://www.xsede.org/software • https://wci.llnl.gov/codes/visit/about.html • http://www.datapreparator.com/ • http://www.jmp.com/software/jmp10/ • http://vis.stanford.edu/wrangler/
Actors	• XSEDE Portal visitor, which in this context is a person interested in using XSEDE for Data Analytics (current or prospective users) • XSEDE Service Providers • XSEDE Portal developers and content managers for documentation • External software developers, e.g., the HDF Group, in collaboration with XSEDE content managers
Prerequisites (Dependencies) and Assumptions)	• XSEDE Portal visitor has Internet access. • XSEDE Portal is available • Active Allocations on XSEDE resources

	● Interactive Resources for Visualization and data manipulation ● Active Allocation for long term storage resource ● Relevant software is installed, tested, documented, and available, including current and relevant legacy versions. This list is open-ended and must be re-evaluated on an ongoing basis, pursuant to user requirements. Software should include visualization tools (ViSIT, Paraview), graphical visualization and analysis tools (KDT) mathematical/statistical tools (R, SAS enterprise, Matlab, SSPS), query tools (MySQL, PostgreSQL), Scripting/Programming Languages (Perl, Python,C++,Python), libraries (NetCDF, HDF). ● User support in the form of contacts and documentation. ● User is authenticated.
Steps	1. Resource Access: The XSEDE Portal visitor (user) logs into the XSEDE computational resource(s) on which the desired data resides. The data may exist as files, in many formats (RDBMS databases, HDFS, GFFS, XWFs, etc.), as files to be loaded into or out of databases (JDBC, ODBC, etc.), as resources accessible at other locations on the Internet, etc. 2. Data Acquisition/Selection: The user fetches data for processing. 3. Data Preparation: XSEDE User uses XSEDE resources to clean, filter and organize collected data for successful mining, modeling and/or analysis, to solve or avoid problems in the data, and presenting the data to the modeling or analysis schema in an optimal way. The user attaches appropriate metadata and provenance information to the resulting data. a. User may specify a "package" they want to run, specify input files (of many possible types, perhaps in different locations, and locality constraints), and the output (file, socket, database, vis tool). b. User may specify general workflows where the "vertices" are the things in (a). This process may be iterative with a human in the loop 4. Data Preservation: The user saves the resulting data, metadata, and provenance information to appropriate long-term storage. Note: intermediate storage requirements may be significant, and output will often be inputs to subsequent stages
Variations (optional)	

Quality Attributes	● **Accuracy**: Infrastructure software must be installed and tested. Documentation must be accurate and accessible, including clear examples. ● **Completeness**: Software to support data gathering, manipulation, and analysis must cover users' needs and community standards for data exchange and reuse. Current versions must be available as well as legacy versions for cases where backward compatibility is not ensured. ● **Context**: Software supporting data gathering, manipulation, and organization should be specified for its relevance to the XSEDE user community. This may imply that some tools, algorithms, and information that are part of the larger data analytics space, for example from the business intelligence community, are not applicable, and therefore not available in XSEDE. ● The owners of the data need to be able to determine who can read and write their data.
Non-functional (optional)	
Issues	● The suite of software required to support data gathering, manipulation, and organization is open-ended and must expand to accommodate users' requirements and evolving community practices. ● Data Preparation Resource should allow for interactive sessions.



j

Data preparation is an iterative process. Users should be able to import data, process data, and evaluate data at any time in the process.

UCDA 3.0	Instrument Data Analysis
Description	User performs on-demand or scheduled analysis of data that is collected from external data sources. Representative data sources include: radio, microwave and optical telescopes, genome sequencers, satellites, electron microscopes, sensor networks, and websites
References	• http://www.xsede.org/resources/overview • http://www.xsede.org/software
Actors	• XSEDE Portal visitor, which in this context is a person interested in using XSEDE for Data Analytics (current or prospective users) • XSEDE Service Providers • XSEDE Portal developers and content managers for documentation • External software developers and instrument developers in collaboration with XSEDE content managers
Prerequisites (Dependencies) and Assumptions)	• XSEDE Portal visitor has Internet access. • XSEDE Portal is available • Active Allocations on XSEDE resources • Active Allocation for long term storage resource • Availability of high speed data collection methods • Data Preparation Ability (see UCDA 2.0) • High throughput data mover nodes and associated software. • Relevant software is installed, tested, documented, and available, including current and relevant legacy versions visualization and data mining and statistical tools. Software should include R, Python, and Hadoop. This list is open-ended and must be re-evaluated on an ongoing basis, pursuant to users' requirements. • User support in the form of contacts and documentation.

Steps	1. **Resource Access:** The XSEDE Portal visitor (user) logs into the XSEDE computational resource(s) on which the desired analytics software resides. 2. **Data Acquisition/Selection**: The user fetches data. The data may exist as files, in databases, as files to be loaded into databases, as resources accessible at other locations on the Internet, etc. 3. **Data Preparation:** UCDA2.0 4. **Use Data:** Run analysis or data mining techniques on data 5. **Data preservation:** Save resulting models, outputs to appropriate long term storage
Variations (optional)	
Quality Attributes	● **Accuracy**: Infrastructure software must be installed and tested. Documentation must be accurate and accessible, including clear examples. ● **Completeness**: Software to support data mining and analysis must cover users' needs and community standards for data exchange and reuse. Current versions must be available as well as legacy versions for cases where backward compatibility is not ensured. ● **Context**: Software supporting data mining and analysis, and should be specified for its relevance to the XSEDE user community. This may imply that some tools, algorithms, and information, which are part of the larger data analytics space, for example from the business intelligence community, are not applicable, and therefore not available in XSEDE.
Non-functional (optional)	
Issues	The suite of software required to support data mining and analysis techniques is open-ended and must expand to accommodate users' requirements and evolving community practices.

Domain
Pre-compiled DatasetExternal Instrument
Tools
Visualiztion RHadoop Python
Algorithms
Frequency
On demandAt RequestScheduled (daily, weekly, hourly)

UCDA 4.0	HPC Simulation Data Analysis
Description	Users analyze data that is obtained from or generated by HPC simulations. The use case encompasses offline analysis of data previously generated by simulation and *in situ* visualization.
References	• http://www.xsede.org/resources/overview • http://www.xsede.org/software
Actors	• XSEDE Portal visitor, which in this context is a person interested in using XSEDE for Data Analytics (current or prospective users) • XSEDE Service Providers • XSEDE Portal developers and content managers for documentation
Prerequisites (Dependencies) and Assumptions)	• XSEDE Portal visitor has Internet access. • XSEDE Portal is available • Active Allocations on XSEDE resources • Active Allocation for long term storage resource • Availability of high speed data collection methods • Data Preparation Ability (see UCDA 2.0) • Relevant software is installed, tested, documented, and available, including current and relevant legacy versions visualization and data mining and statistical tools. This list is open-ended and must be re-evaluated on an ongoing basis, pursuant to user requirements. • User support in the form of contacts and documentation.
Steps	1. Resource Access: The XSEDE Portal visitor (user) logs into the XSEDE computational resource(s) on which the desired analytics software resides. 2. Data Acquisition/Selection (see variations below) 3. Data Preparation (see variations below) 4. Use Data: Run analysis or data mining techniques on data 5. Data preservation: Save resulting models, outputs to appropriate long term storage
Variations (optional)	Step 2: • Post Processing: the user fetches data. The data may exist as files, in databases, as files to be loaded into databases, as resources accessible at other locations on the Internet, etc.

	• In Situ: Data is collected in situ on same resource as analysis is being run Step 3: • Post Processing: Will require Data preparation as described in UCDA2.0 • In Situ: Will manage data preparation in Situ
Quality Attributes	• **Accuracy**: Applications and resources available via XSEDE changes over time and documentation will need to be updated. Infrastructure software must be installed and tested. Documentation must be accurate and accessible, including clear examples. • **Completeness**: Software to support dynamic workflow analysis and manipulation must cover users' needs and community standards for data exchange and reuse. Current versions must be available as well as legacy versions for cases where backward compatibility is not ensured. • **Context**: Workflow analysis and manipulation resources and information should be defined in terms that are relevant for the XSEDE user community.
Non-functional (optional)	
Issues	The suite of software required to support data mining and analysis techniques is open-ended and must expand to accommodate users' requirements and evolving community practices. Ensure in situ activities do not adversely affect other users.

UCDA 5.0	In-situ Computational Steering
Description	Users performs computational steering, i.e. views and/or visualizes the results of a simulation while it is running and subsequently takes action, such as changing simulation parameters.
References	• http://www.xsede.org/resources/overview • http://www.xsede.org/software
Actors	• XSEDE Portal visitor, which in this context is a person interested in using XSEDE for Data Analytics (current or prospective users) • XSEDE Service Providers • External software developers, in collaboration with XSEDE content managers
Prerequisites (Dependencies) and Assumptions)	• Active XSEDE Portal Access • Active Allocations on XSEDE resources • Interactive Resources to run small scale (fast) visualization or analysis • Interactive, On-Demand resources available if required for diagnostic evaluation
Steps	1. The XSEDE Portal visitor (user) logs into the XSEDE computational resource(s) on which to run. 2. XSEDE user modifies code as necessary to read from parameters file in situ to allow computational steering 3. XSEDE user creates parameter file(s). Parameter files can include parameters that cause the program to dump diagnostics for user to check at certain intervals. 4. XSEDE user launches job. 5. XSEDE User HPC at request resources, interactive computing capabilities, to evaluate simulation diagnostics, or other intermediate output 6. As needed user has ability and tools to modify parameter file(s) in situ as response to diagnostics evaluation

Variations (optional)	
Quality Attributes	● **Accuracy**: Applications and resources available via XSEDE changes over time and documentation will need to be updated. Infrastructure software must be installed and tested. Documentation must be accurate and accessible, including clear examples. ● **Completeness**: Software to support dynamic workflow analysis and manipulation must cover users' needs and community standards for data exchange and reuse. Current versions must be available as well as legacy versions for cases where backward compatibility is not ensured. ● **Context**: Workflow analysis and manipulation resources and information should be defined in terms that are relevant for the XSEDE user community.
Non-functional (optional)	
Issues	● Allowing users to manipulate runtime parameters should not have negative impact on resources performance or queues. ● The suite of software required to support data mining and analysis techniques is open-ended and must expand to accommodate users' requirements and evolving community practices.